# Towards Streaming Speech Translation for Real-world Scenarios

**福田 りょう**

奈良先端科学技術大学院大学 中村研究室

（現在：NTTコミュニケーション科学基礎研究所）

2024.06.20

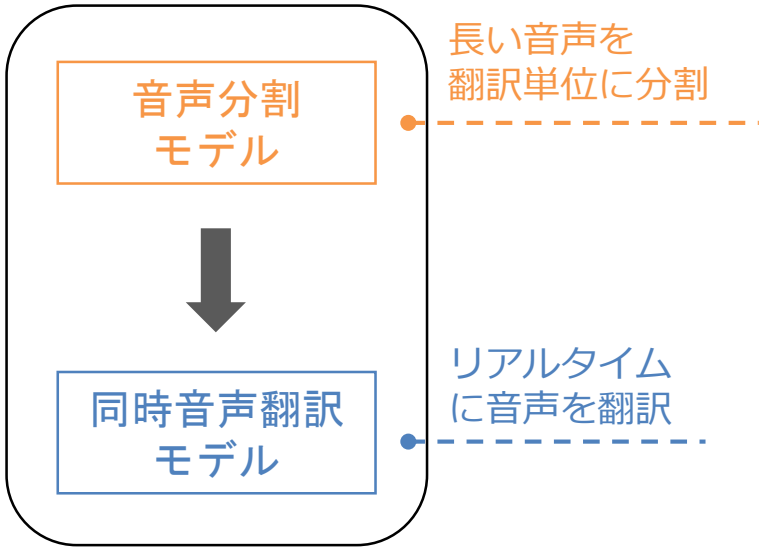
第11回AAMT長尾賞学生奨励賞招待講演

# 研究の概要

長い音声をリアルタイムに翻訳する音声翻訳システムの構築

- 講演など、一人の話者が長時間（数分～）話しているような場面
- 音声を固定長のフレームで継続的に受け取りつつ、リアルタイムに翻訳

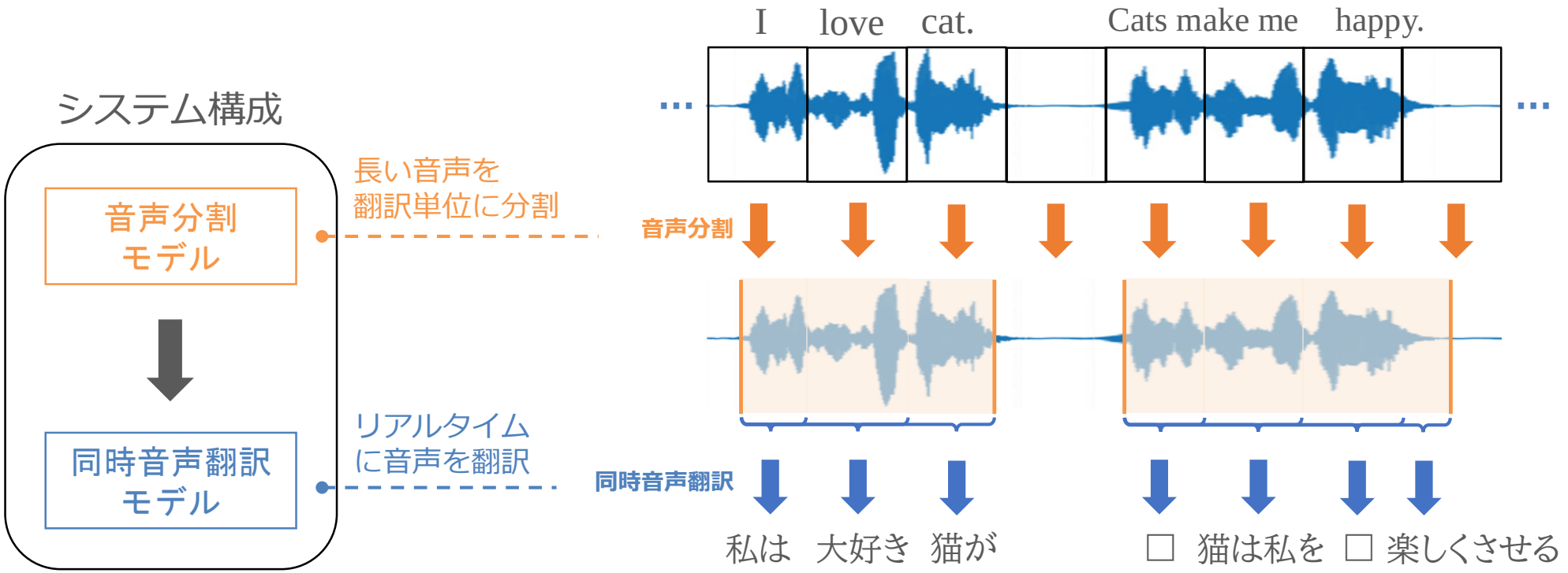
システム構成



# 研究の概要

長い音声をリアルタイムに翻訳する音声翻訳システムの構築

- 講演など、一人の話者が長時間（数分～）話しているような場面
- 音声を固定長のフレームで継続的に受け取りつつ、リアルタイムに翻訳



# 発表内容

1. Speech Segmentation for Long-form Speech Translation
   - ① 音声分割 に関する取り組み

2. Streaming Simultaneous Speech Translation
   - ② 同時音声翻訳 に関する取り組み
   - ①と②を組み合わせたシステム構築

# 発表内容

1. Speech Segmentation for Long-form Speech Translation
   - ① 音声分割 に関する取り組み

2. Streaming Simultaneous Speech Translation
   - ② 同時音声翻訳 に関する取り組み
   - ①と②を組み合わせたシステム構築

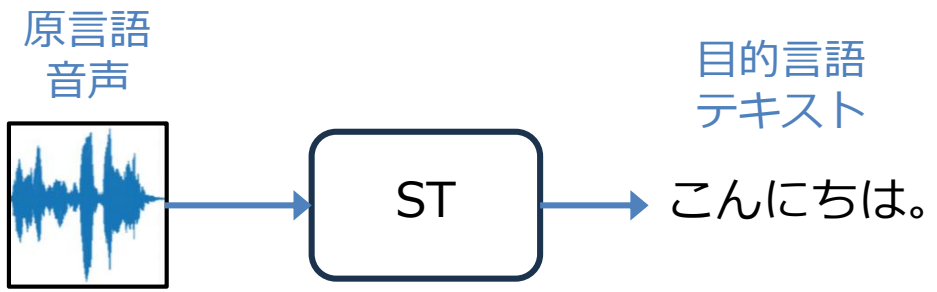
# 音声機械翻訳 (Speech Translation; ST)

**ST**：音声を入力とする機械翻訳（MT）

- 出力はテキスト/音声 … 今回はテキスト出力を想定
- ニューラル機械翻訳 (NMT) に基づくST：**Cascade ST**と**End-to-end ST**

**Cascade ST**
[WM Stentiford and G Steer, 1988]


**End-to-end ST**
[Bérard+, 2016]

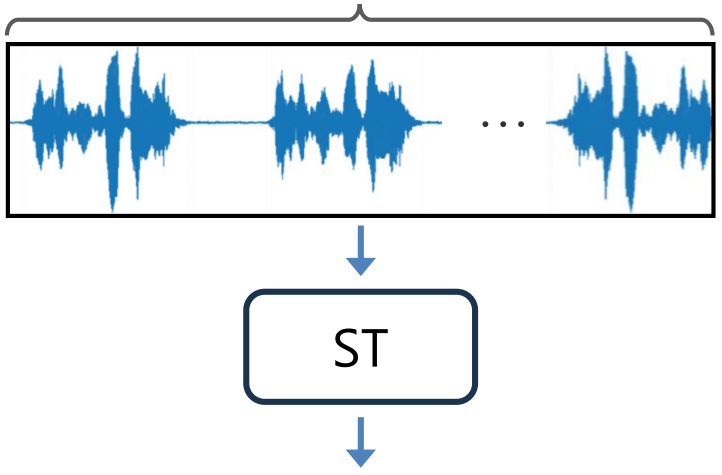


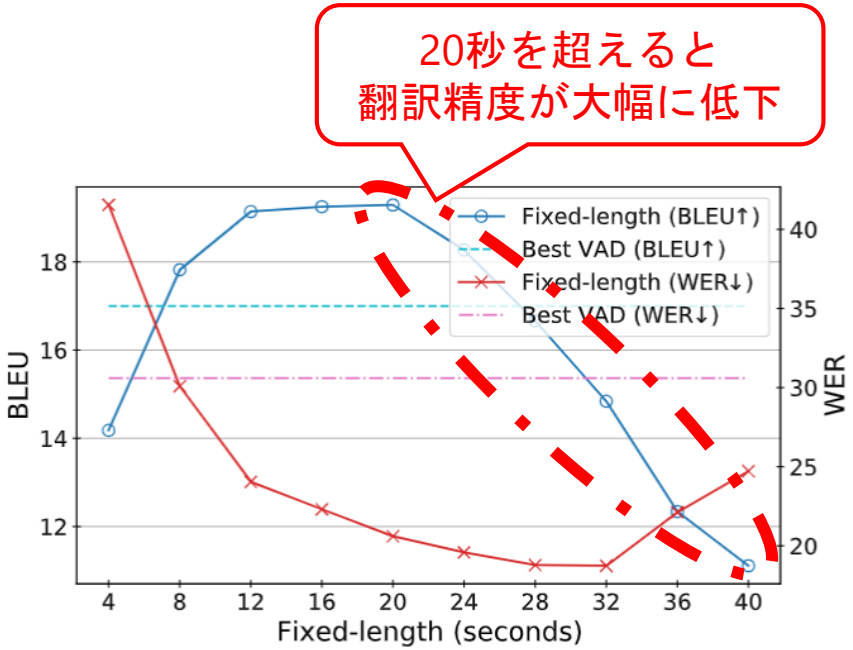
# STのための自動音声分割

## 長い音声の翻訳は難しい

- テキスト入力のMTに比べて入力が非常に長い
- 入力が長くなると … 低品質な翻訳 / 生成失敗
  - 原因：計算量の増加 & 学習時とのギャップ

10分の音声 -> 系列長 9,600,000フレーム



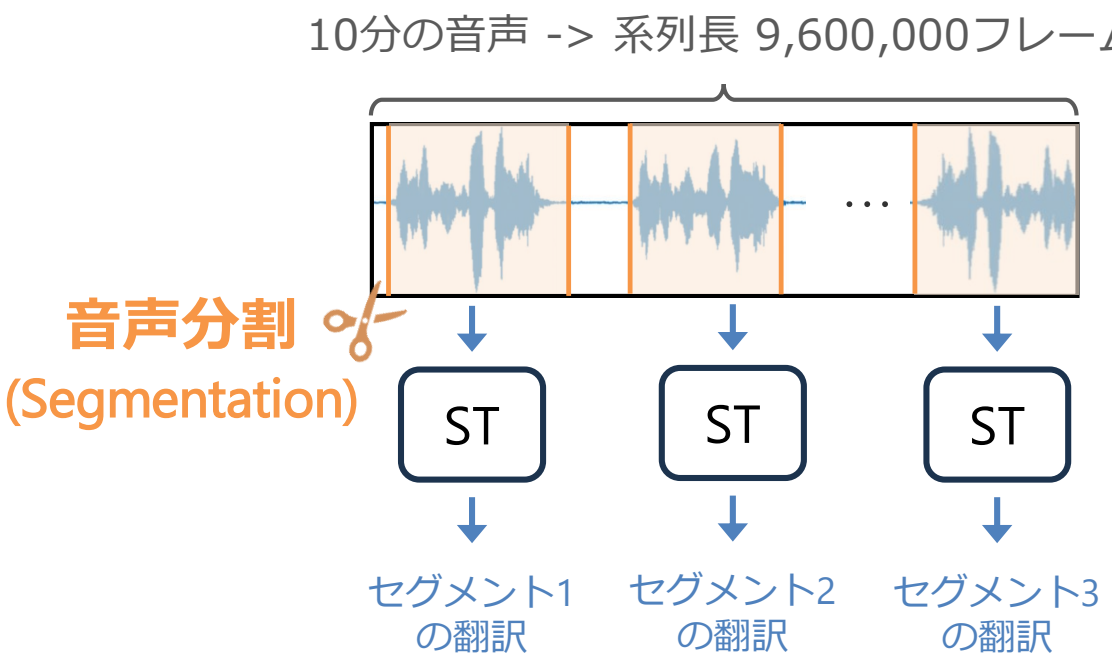



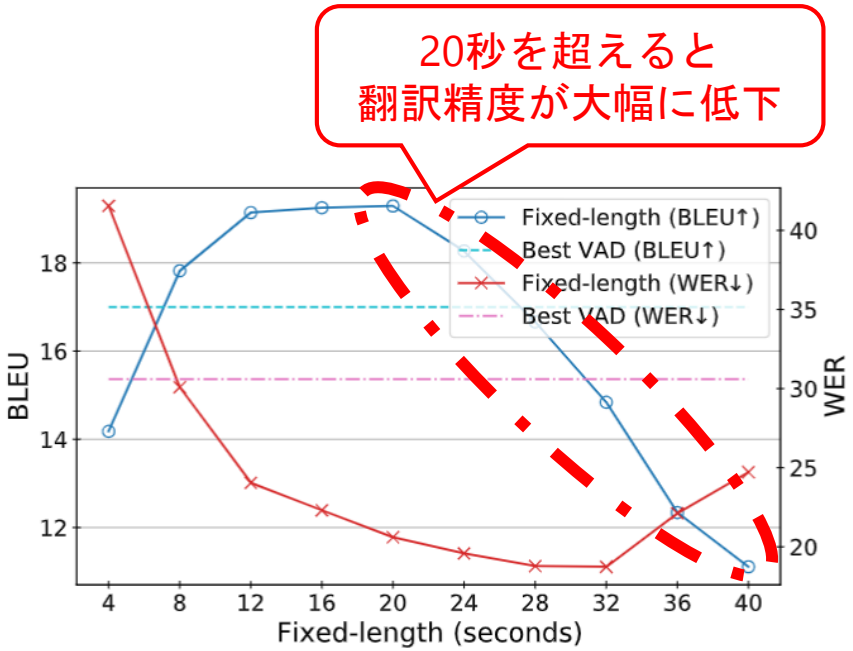

[Fukuda+, 2022]

# STのための自動音声分割

**長い音声の翻訳は難しい ⇒ 音声分割 (Segmentation) が必要**

- テキスト入力のMTに比べて入力が非常に長い
- 入力が長くなると … 低品質な翻訳 / 生成失敗
  - 原因：計算量の増加 & 学習時とのギャップ




[Fukuda+, 2022]

# 音声分割の従来手法：VADによる分割

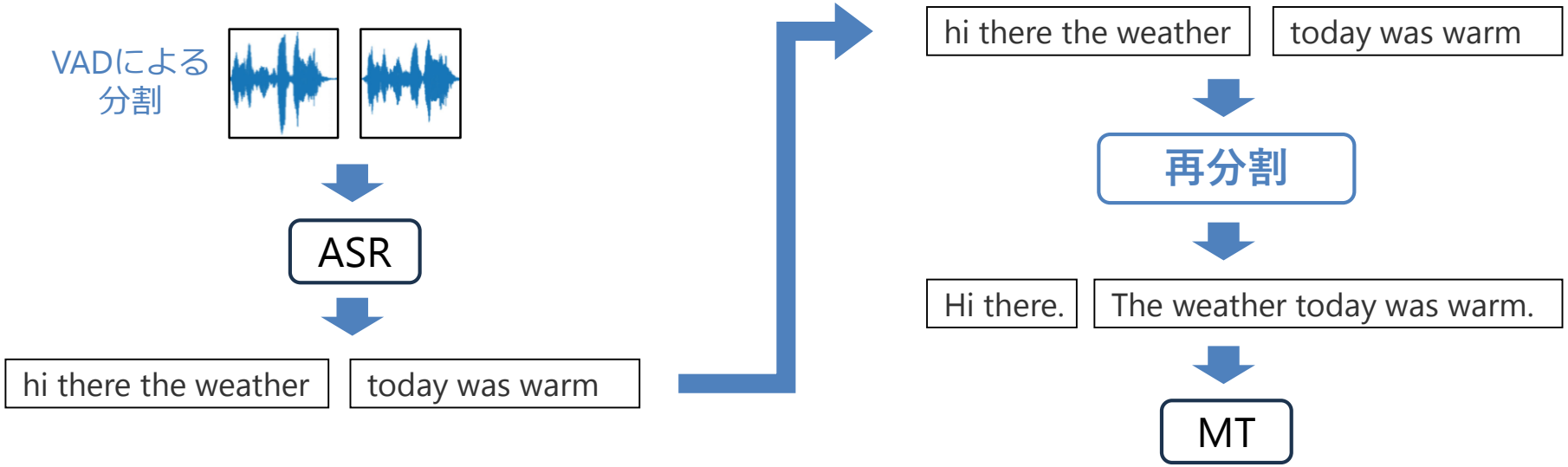
音声区間検出 (**VAD**; Voice Activity Detection) [Sohn+99][Bangalore+12]

- 検出した音声区間に基づいて音声を分割
- **無音は必ずしも文の境界と一致しない** → 過剰な分割による翻訳精度の低下

# 音声分割の従来手法：テキスト再分割

句読点予測モデル [Niehues+2015] や言語モデル [Stolcke and Shriberg, 1996] などを用いて音声認識結果を**文単位**に再分割

- Cascade STにおける有用性が広く知られている
  - 文単位のセグメントは翻訳に適している
- 中間の音声認識結果が必要で、End-to-end STに向かない



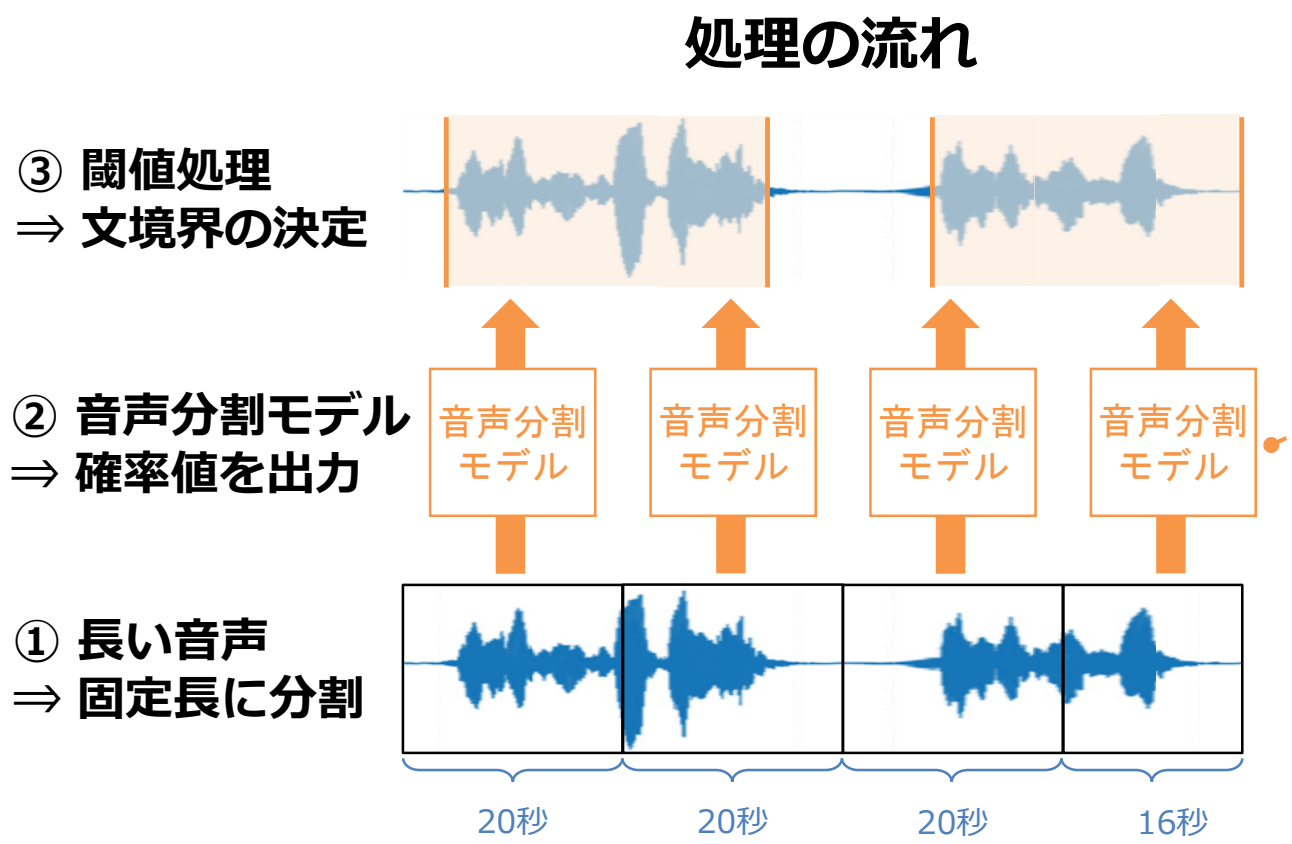
# 提案手法

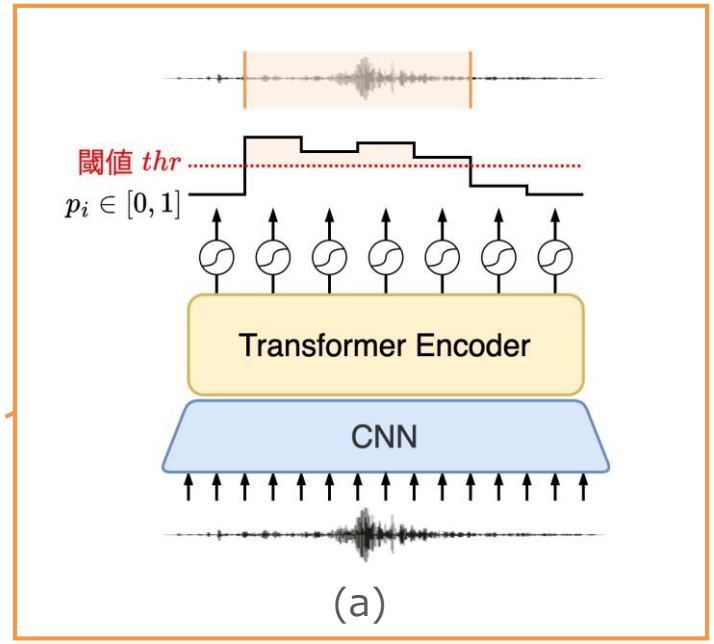
**文の境界**を**音声**から直接予測する音声分割モデル

| 分割手法 | 分割単位 | 入力 | 利点 |
|---|---|---|---|
| VAD | 音声区間 | 音声 | End-to-end STに利用しやすい |
| テキスト分割 | 文の境界 | テキスト | 翻訳精度が高い |
| **本研究** | 文の境界 | 音声 | **End-to-end STに利用しやすく<br>翻訳精度が高い** |

# 提案手法

## 文の境界を音声から直接予測する音声分割モデル



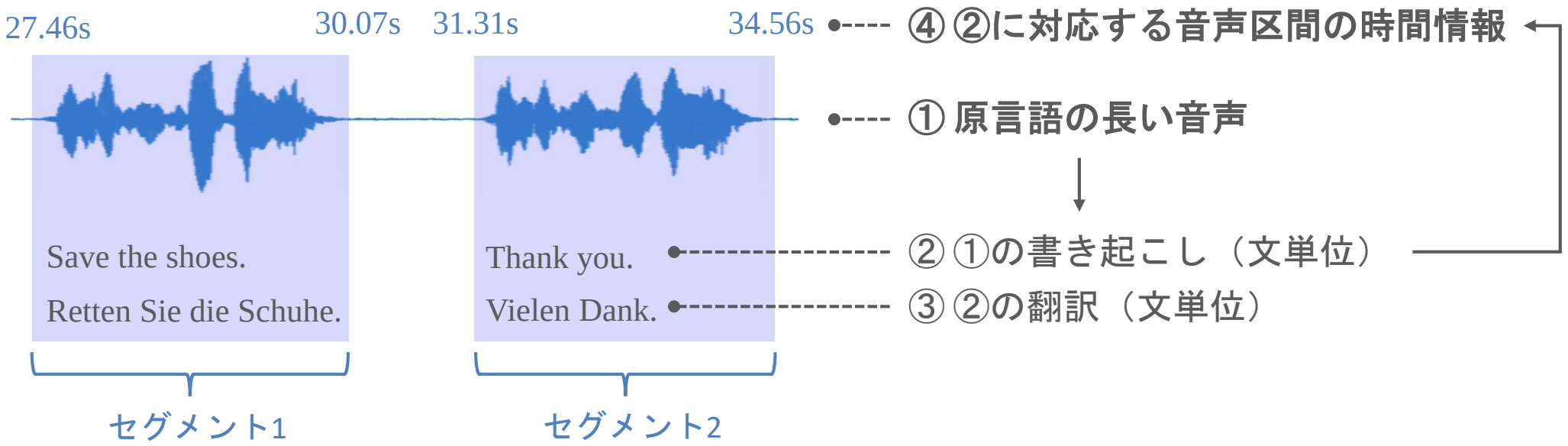

モデル構造
(a) Transformer [Vaswani+, 2017] に基づくモデル [Fukuda+, 2022]
(b) wav2vec 2.0 [Baevski+, 2020] に基づくモデル [Fukuda+, 2023a]

Skip

# 提案手法

## 文の境界を音声から直接予測する音声分割モデル

- 音声翻訳コーパスの文境界を学習
  - ①から④を予測するタスク

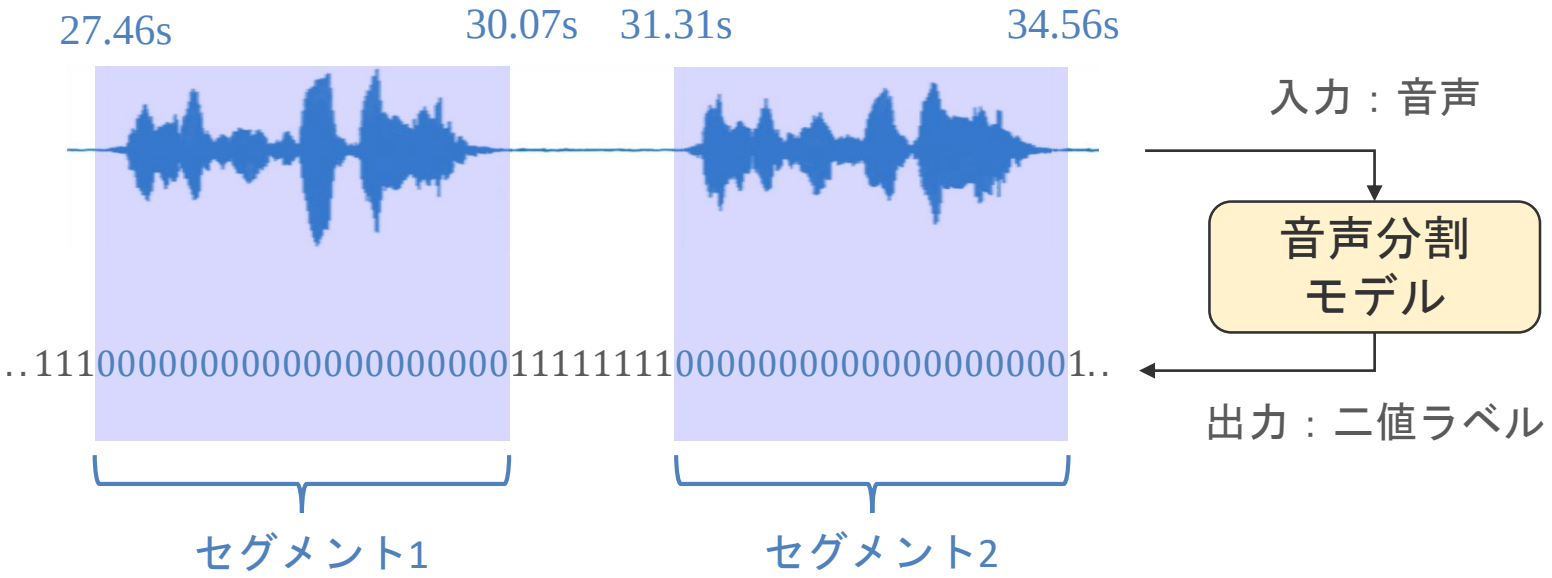


Skip

# 提案手法

## 文の境界を音声から直接予測する音声分割モデル

- 音声翻訳コーパスの文境界を学習
  - ①から④を予測するタスク
- 音声フレームを0 (文の中) / 1 (文の外) に分類する系列ラベリング



# 実験：音声分割手法の比較

TED talksの英語講演音声 ⇒ ドイツ語、日本語への翻訳実験 ※補足資料に実験設定

- 音声分割モデルは比較手法を上回った
  - 両言語ペア（英独、英日）、両STシステムで一貫した翻訳精度の向上
- VADとの組み合わせで更に翻訳精度が向上した

英語 → ドイツ語

| | | Cascade | End-to-end |
|---|---|---|---|
| | | BLEU↑ | BLEU↑ |
| トップライン | 人手分割 | 23.59 | 22.50 |
| 比較手法 | VAD | 17.02 | 16.40 |
| | Fixed-length | 19.29 | 17.96 |
| 提案手法 | 音声分割モデル | 20.18 | 19.10 |
| | + VAD | **20.99** | **19.87** |

英語 → 日本語

| | Cascade | End-to-end |
|---|---|---|
| | BLEU↑ | BLEU↑ |
| 人手分割 | 12.50 | 10.60 |
| VAD | 9.26 | 8.14 |
| Fixed-length | 9.64 | 8.52 |
| 音声分割モデル | 9.71 | 8.77 |
| + VAD | **10.60** | **9.24** |

# 発表内容

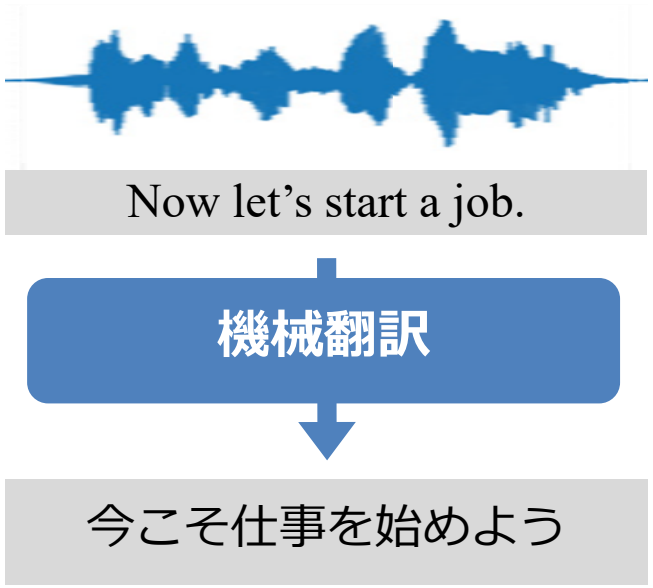
1. Speech Segmentation for Long-form Speech Translation
   - ① 音声分割 に関する取り組み

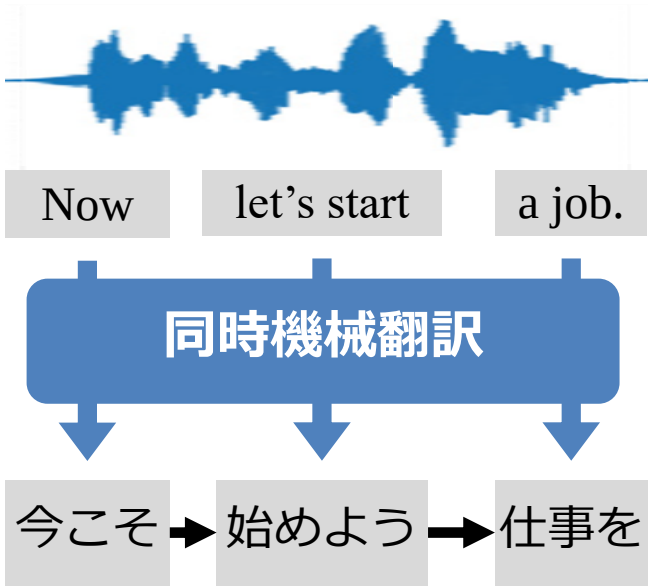
2. Streaming Simultaneous Speech Translation
   - ② 同時音声翻訳 に関する取り組み
   - ①と②を組み合わせたシステム構築

# 同時機械翻訳

リアルタイムに翻訳を行うための機械翻訳技術

- MTやSTモデルが文の終了を待たずに翻訳を開始
- 技術的な難しさ：訳出タイミングの決定、原言語の語順に近い翻訳



# 同時機械翻訳の課題と取り組み

- **性能向上のための課題：モデルの学習データの不足**
  - オフライン翻訳データからの学習には限界（特に、語順が大きく異なる言語対）
  - 同時通訳データの整備 [Tohyama+, 2004][松原+, 2001][松下+, 2020][Doi+, 2021]
  - 同時通訳データによるMTの学習 [Ryu+, 2004][Shimizu+, 2013][Ko+, 2023]
  - 順送り訳データの作成 [福田+, 2024a]

# 同時機械翻訳の課題と取り組み

- **実用上の課題：長い音声の翻訳が難しい**
  - オフラインSTと同様、翻訳モデル単体で長い音声を処理することは困難
    ⇒ **音声分割の必要性**

  - 実環境でSTシステムを動作させるために考えられる方法
    - **無音に基づく自動分割**：高速な一方、翻訳精度の低下が大きい
    - **ユーザーが発話終了を明示的に知らせる**：リアルタイムな翻訳には不向き

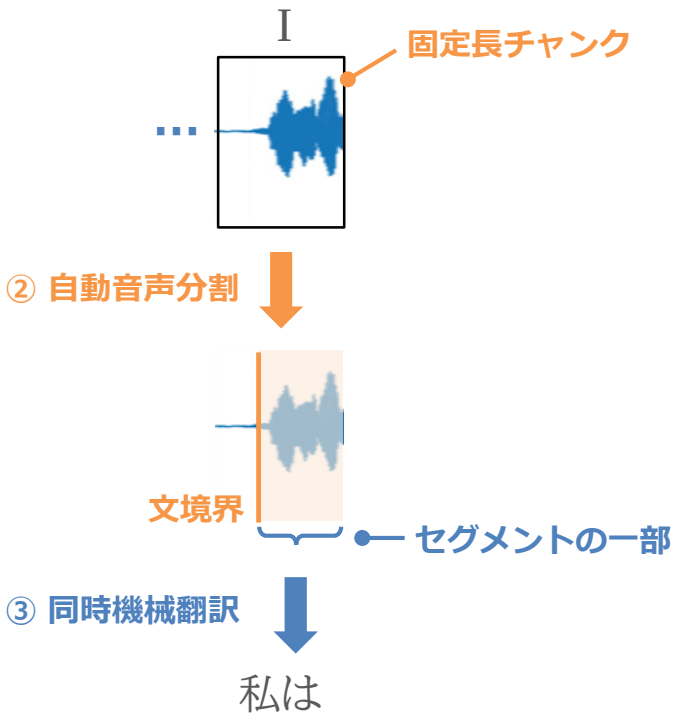
  - 同時機械翻訳の研究のほとんどが長い音声では実験を行っていない
    - 事前分割された短い音声で比較

  - **本研究：**オフラインSTで高精度な音声分割モデルを同時機械翻訳に適用 [福田+, 2024b]

# ストリーミングSTの構築

長い音声ストリームに追従するオンライン翻訳システム

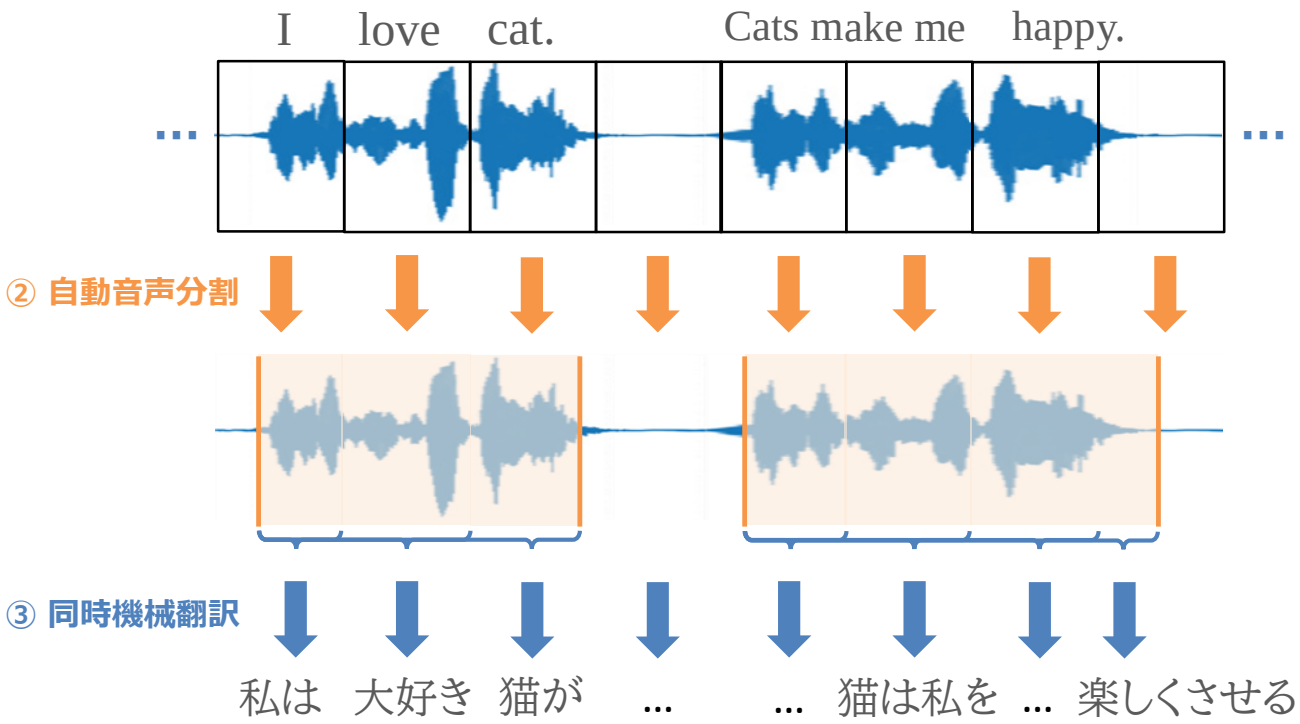
- 音声分割モデルと同時機械翻訳モデルが交互に動作
- TED talks などの長い音声をリアルタイムに翻訳できる



# ストリーミングSTの構築

長い音声ストリームに追従するオンライン翻訳システム

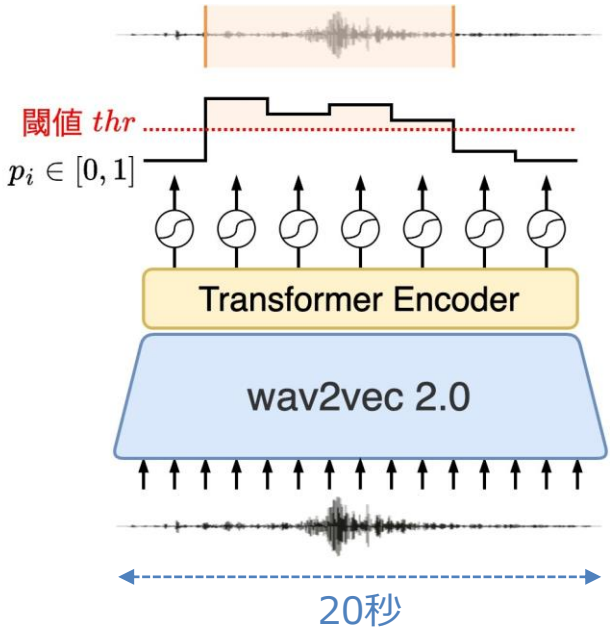
- 音声分割モデルと同時機械翻訳モデルが交互に動作
- TED talks などの長い音声をリアルタイムに翻訳できる



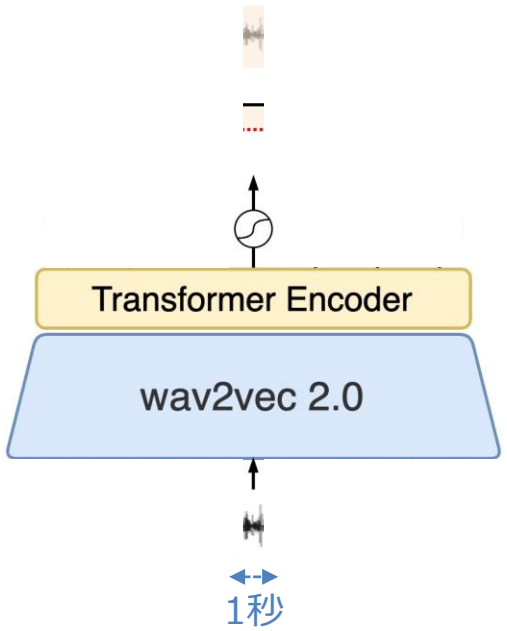
# ストリーミングST：音声分割モデル

Wav2vec 2.0に基づく音声分割モデル [Fukuda+, 2023a] を同時機械翻訳に適用

- オフラインSTでは長い音声チャンク（20秒）を入力して高精度な予測
- チャレンジ：同時機械翻訳は1秒程度の短いチャンクを入力とするため、少ない情報でセグメント境界を予測する必要がある



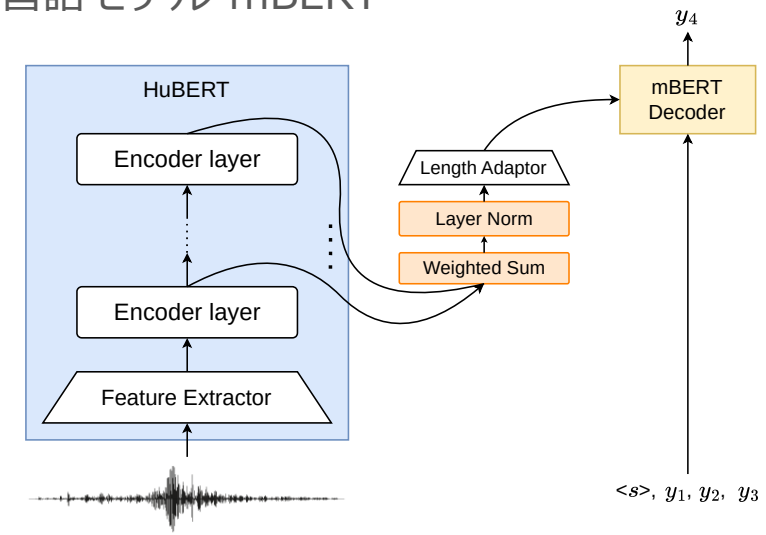

オフラインSTの音声分割




同時機械翻訳の音声分割

# ストリーミングST：同時機械翻訳モデル

- モデルの構造は通常のSTと同様
  - Transformerに基づくEncoder-Decoderモデル
  - 英語音声モデル HuBERT + 多言語モデル mBERT




[Fukuda+, 2023b]

- 通常のSTモデルで同時翻訳を行うための工夫
  - ① **Prefix Alignment** [Kano+, 2022]
  - ② **Local Agreement** [Liu+, 2020]

# ストリーミングST：同時機械翻訳モデル

① **Prefix Alignment**: モデルの学習手法

- オフライン翻訳データでは文の途中から翻訳することを学習できない
- オフライン翻訳データから抽出した**部分的な翻訳ペア（Prefix pair）**で学習

| Partial input | Partial output (*gloss*) | Offline translation |
|---|---|---|
| I | 私は。(*I.*) | 私はペンを買った |
| I bought | 私は買った。 (*I bought*) | |
| I bought a | 私は一つ買った (*I bought one*) | |
| I bought a pen | 私はペンを買った (*I bought a pen*) | |

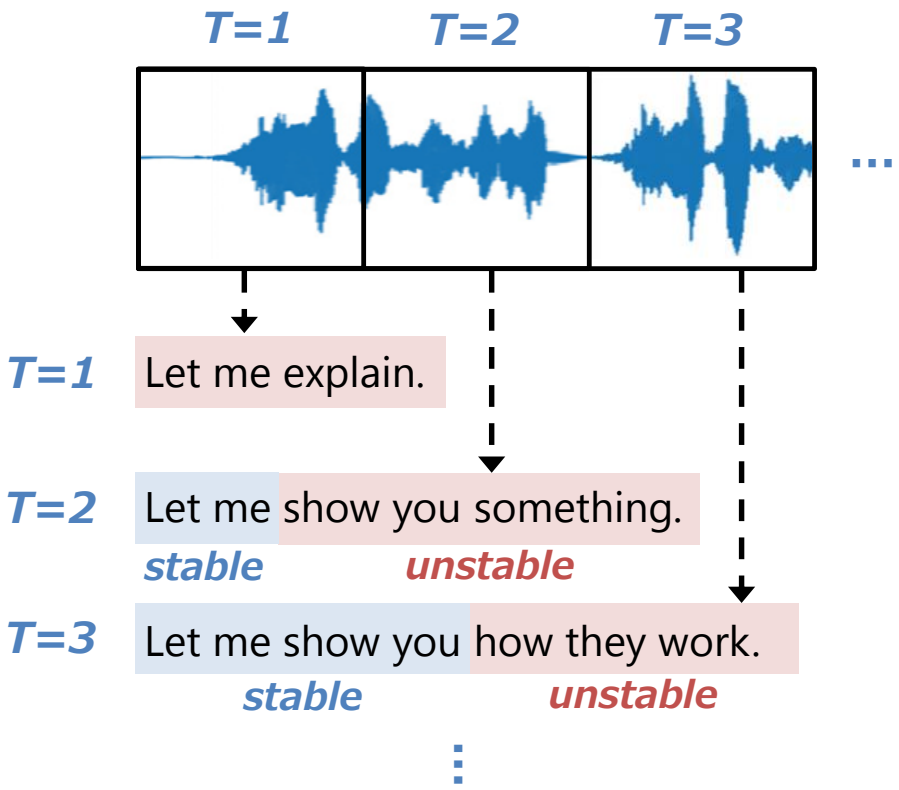
Prefix-to-prefix pairs

[(“I”, “私は”),
(“I bought pens.”, “私はペンを買った。”)]

# ストリーミングST：同時機械翻訳モデル

② **Local Agreement**: モデルの推論方略

- シンプルな方法でREAD/WRITEのタイミングを考える必要がない
- 固定長チャンクを翻訳 ⇒ 前時刻の翻訳との**共通部分**のみ出力として確定

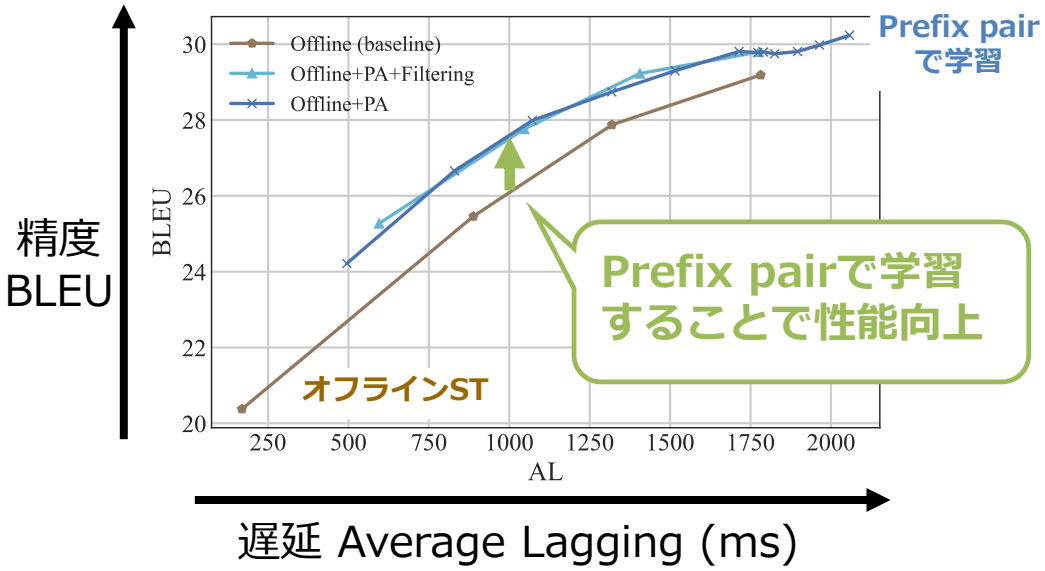


# 実験1: 同時機械翻訳モデルの評価

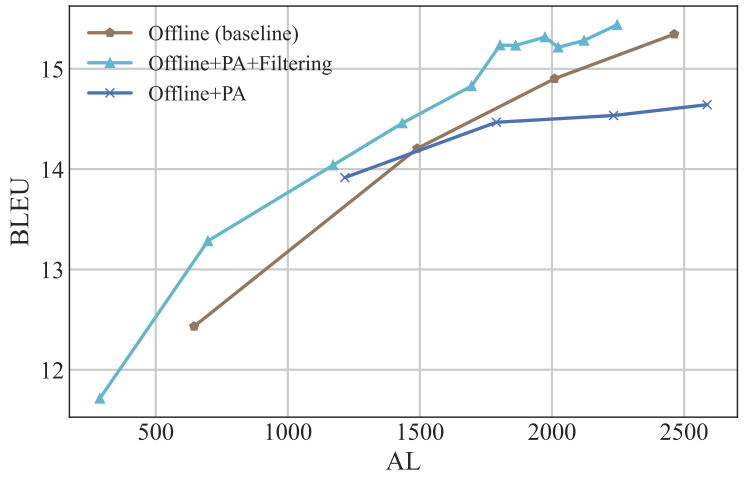
分割済みの短い音声で、同時機械翻訳モデルの性能を評価　※補足資料に実験設定

- TED talksの英語講演音声 ⇒ ドイツ語、日本語、中国語

英語→ドイツ語




英語→日本語




入力チャンクの大きさ
(200~1000ms) で調節

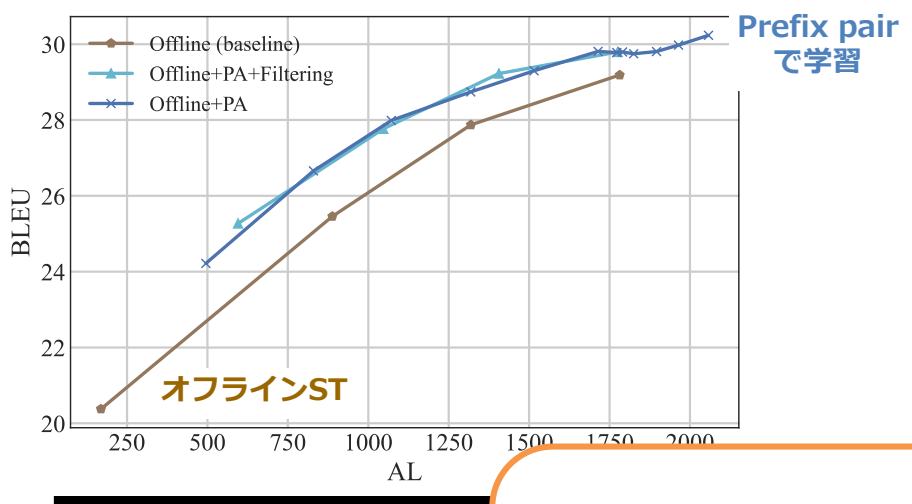
**※実時間遅延は
1~3秒程度**

# 実験1: 同時機械翻訳モデルの評価

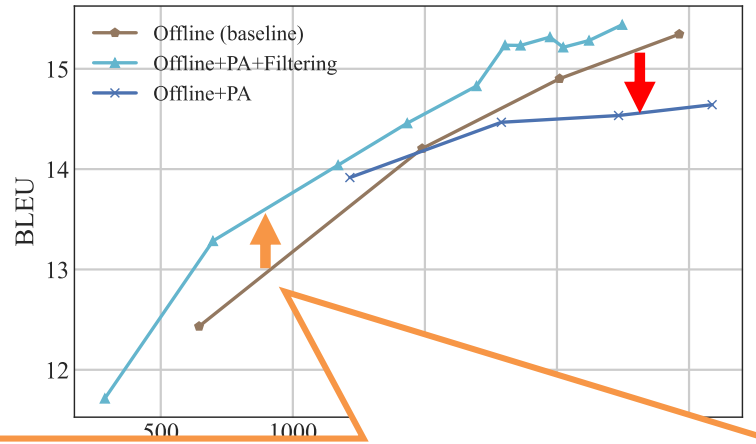
分割済みの短い音声で、同時機械翻訳モデルの性能を評価　※補足資料に実験設定

- TED talksの英語講演音声 ⇒ ドイツ語、日本語、中国語

英語→ドイツ語

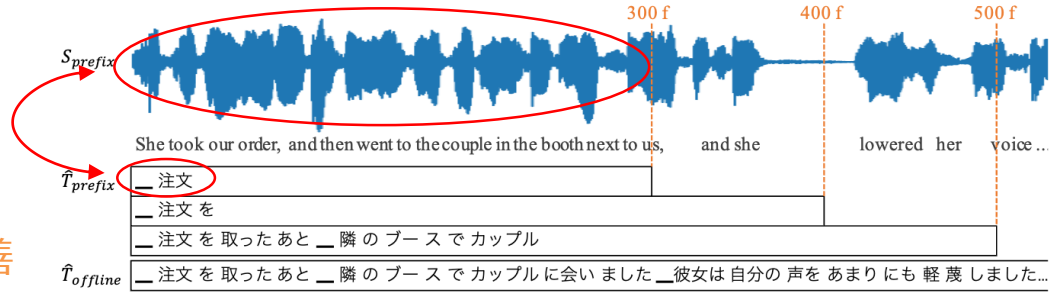



英語→日本語




英日は語順の違いから
低品質なPrefix pairが
多く、性能悪化

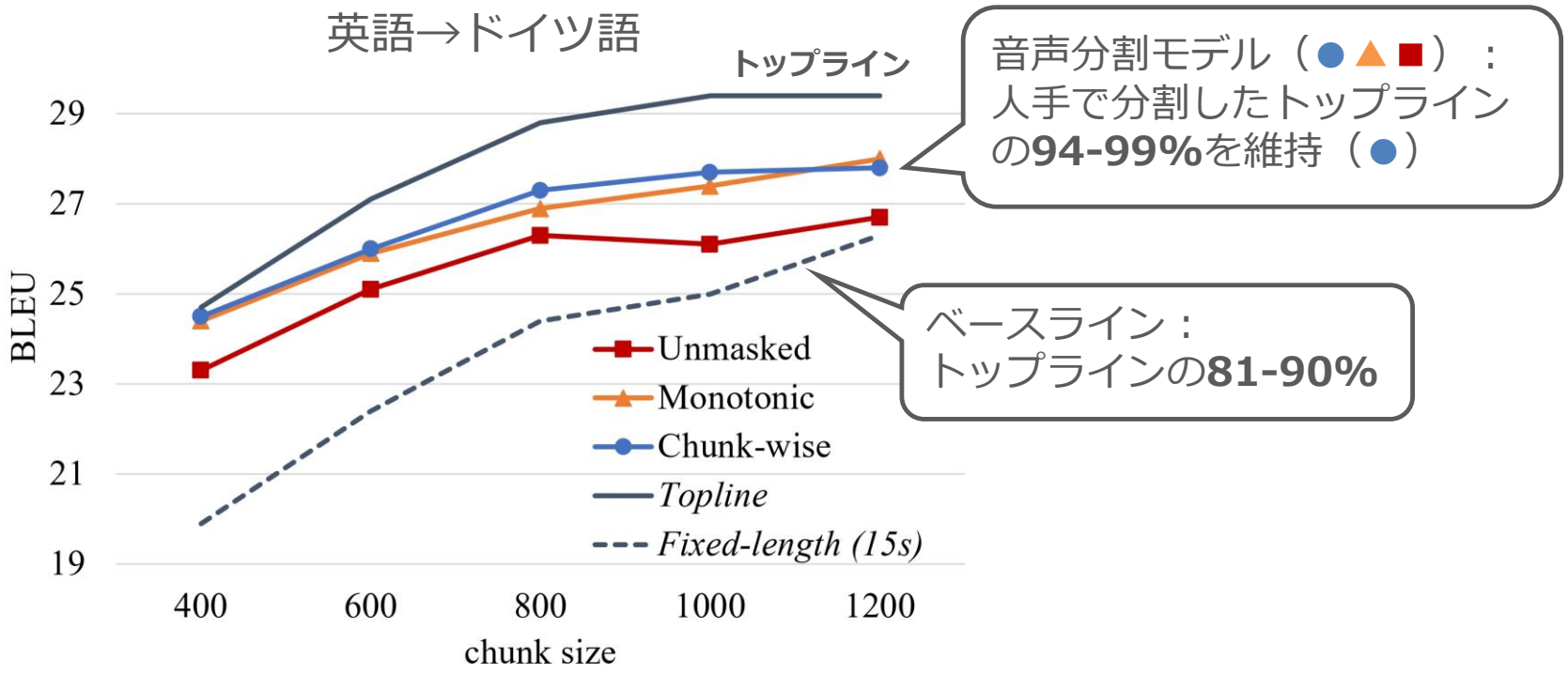
⇒ フィルタリングで改善



# 実験2: ストリーミングSTシステムの評価

長い音声を同時翻訳するストリーミングSTで、音声分割手法を比較　※補足資料に実験設定

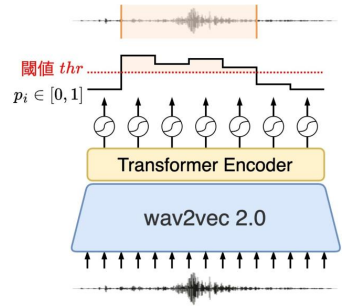
- 音声分割モデルはトップラインの翻訳精度を94-99%維持 ⇒ 同時機械翻訳での有用性



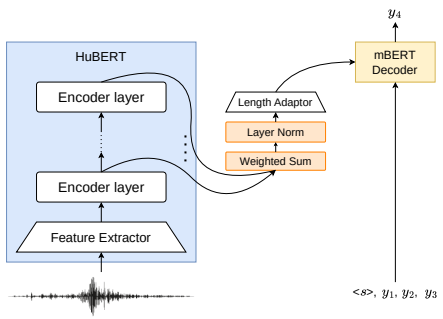
# まとめ

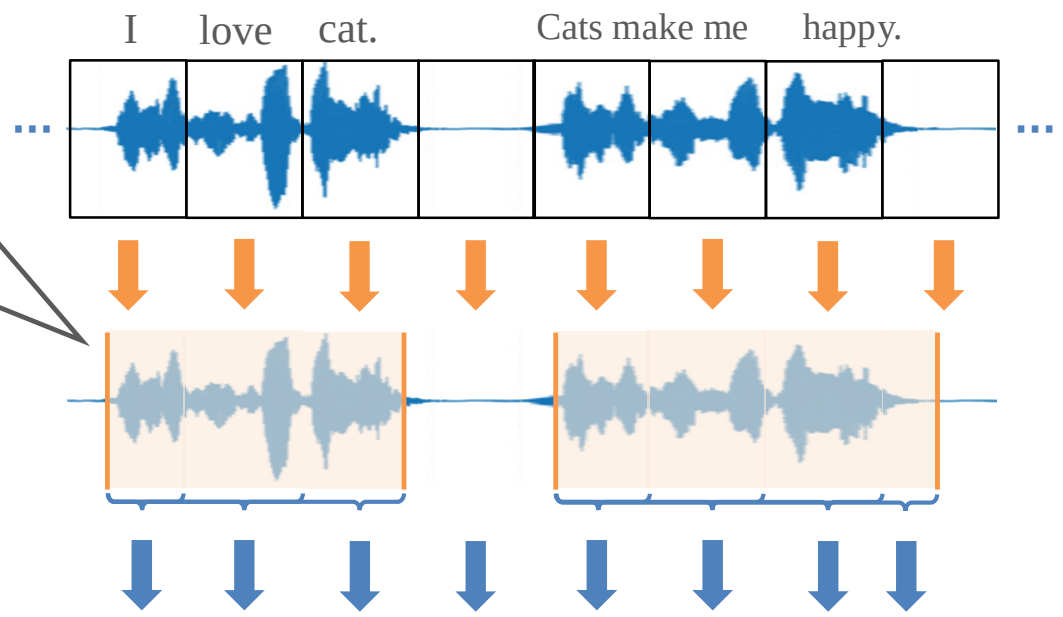
長い音声ストリームに追従するオンライン翻訳システムの構築

(1) 自動音声分割の性能向上




(2) 同時機械翻訳の性能向上




(3) モデル統合
・システム評価

# 今後の展望

- 同時機械翻訳の性能向上
  - 語順の違いが大きい言語ペアではまだ低遅延・高精度の両立は難しい
  - 同時通訳データ等の活用により、同時通訳者の技能（順送り訳、省略）を再現
- 処理速度の改善
  - 音声分割モデル・同時機械翻訳モデルともに計算効率に改善の余地が大きい
  - 音声分割を内部で行う機械翻訳モデルなども有望な方向性
- 実環境にロバストなシステム
  - 自発的な話し言葉を処理
  - 騒がしい環境・複数人が話しているような状況での処理

# 文献


**音声分割**

- Ryo Fukuda, Katsuhito Sudoh and Satoshi Nakamura. Improving Speech Translation Accuracy and Time Efficiency with Fine-tuned wav2vec 2.0-based Speech Segmentation. IEEE/ACM Transactions on Audio, Speech, and Language Processing, doi: 10.1109/TASLP.2023.3343614. [Fukuda+, 2023a]
- Ryo Fukuda, Katsuhito Sudoh and Satoshi Nakamura. Speech Segmentation Optimization using Segmented Bilingual Speech Corpus for End-to-end Speech Translation. In Proceedings of Interspeech 2022, pp. 121–125. Incheon, Korea, Oct. 2022. [Fukuda+, 2022]

**同時翻訳**

- 福田りょう, 土肥康輔, 須藤克仁, 中村哲. 原発話に忠実な英日同時機械翻訳の実現に向けた順送り訳評価データ作成. 第259回自然言語処理研究発表会, 兵庫, 3月, 2024. [福田+, 2024a]
- 福田りょう, 須藤克仁, 中村哲. 漸進的な音声分割を用いたストリーミング同時音声翻訳. 言語処理学会 第30回 年次大会 (NLP 2024), 兵庫, 3月, 2024. [福田+, 2024b]
- Ryo Fukuda, Yuta Nishikawa, Yasumasa Kano, Yuka Ko, Tomoya Yanagita, Kosuke Doi, Mana Makinae, Sakriani Sakti, Katsuhito Sudoh and Satoshi Nakamura. NAIST Simultaneous Speech Translation System for IWSLT 2023. In Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT), pp. 330–340. Toronto, Canada, Jul. 2023. [Fukuda+, 2023b]
- Ryo Fukuda, Sashi Novitasari, Yui Oka, Yasumasa Kano, Yuki Yano, Yuka Ko, Hirotaka Tokuyama, Kosuke Doi, Tomoya Yanagita, Sakriani Sakti, Katsuhito Sudoh and Satoshi Nakamura. Simultaneous Speech-to-speech Translation System with Transformer-based Incremental ASR, MT, and TTS. In Proceedings of the 24th Conference of the Oriental COCOSDA (O-COCOSDA), pp. 186–192. Singapore (Online), Nov. 2021. [paper]

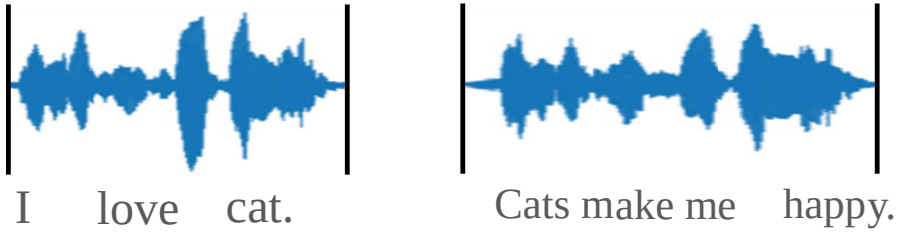
# 補足資料

# 研究の概要

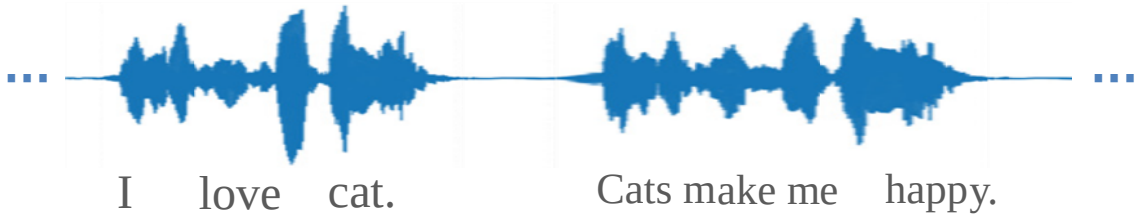
## 長い音声ストリームに追従するオンライン翻訳システムの構築

複数の発話が連続するような
長い音声（⇔短い音声）

発話で分割された
短い音声



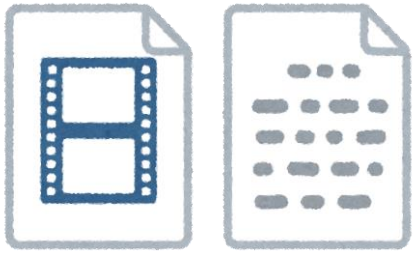

分割されていない
長い音声



# 研究の概要

長い音声ストリームに追従するオンライン翻訳システムの構築

ある程度のリアルタイム性を
持った翻訳（⇔オフライン翻訳）



オフライン翻訳
動画や文書の翻訳

オンライン翻訳
逐次通訳や同時通訳

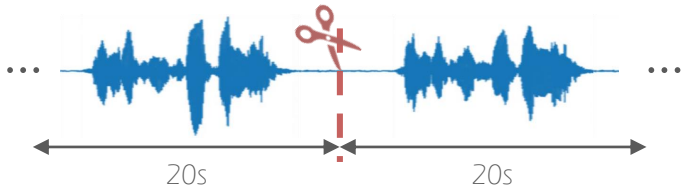
# 発表内容

1. Speech Segmentation for Long-form Speech Translation
   - ① 音声分割 に関する取り組み

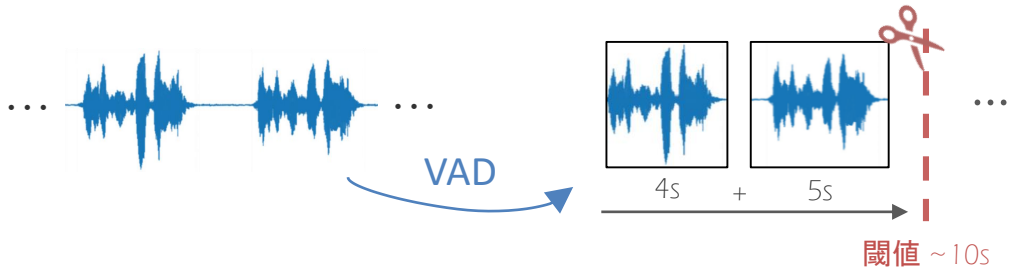
2. Streaming Simultaneous Speech Translation
   - ② 同時音声翻訳 に関する取り組み
   - ①と②を組み合わせたシステム構築

# 音声分割の従来手法：長さに基づく分割

- 固定長分割 [Sinclair+14] [Gaido+21]
  - STが翻訳しやすい長さのセグメントへ分割できる




- 無音に基づく分割と固定長分割の組み合わせ [Potapczyk+20][Gaido+21][Inaguma+21]
  - セグメントの長さを考慮することで、VADによる過剰分割を緩和できる

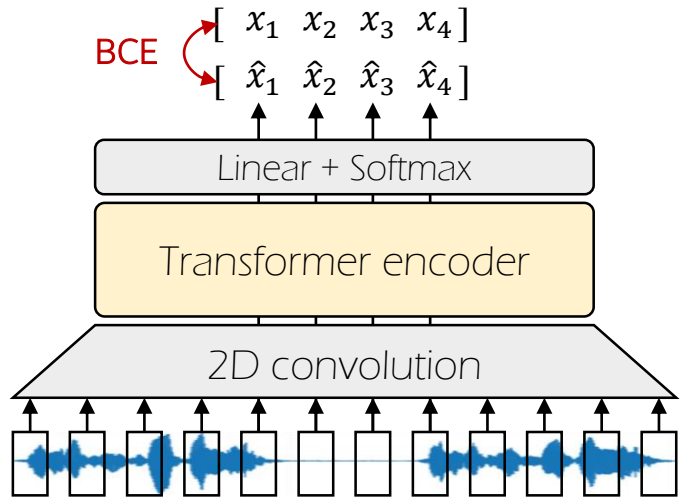



音声や言語の情報を考慮しない
→ 発話を分断するなど、不適切な分割による翻訳精度の低下

# 提案手法：音声分割モデルの構造

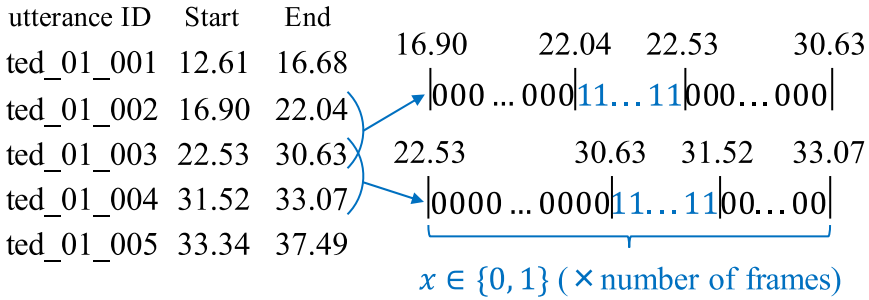
## 発話分割モデル

- 2D convolution + Transformer Encoder
- 入力：83次元の音響特徴量（FBANK+ピッチ情報）



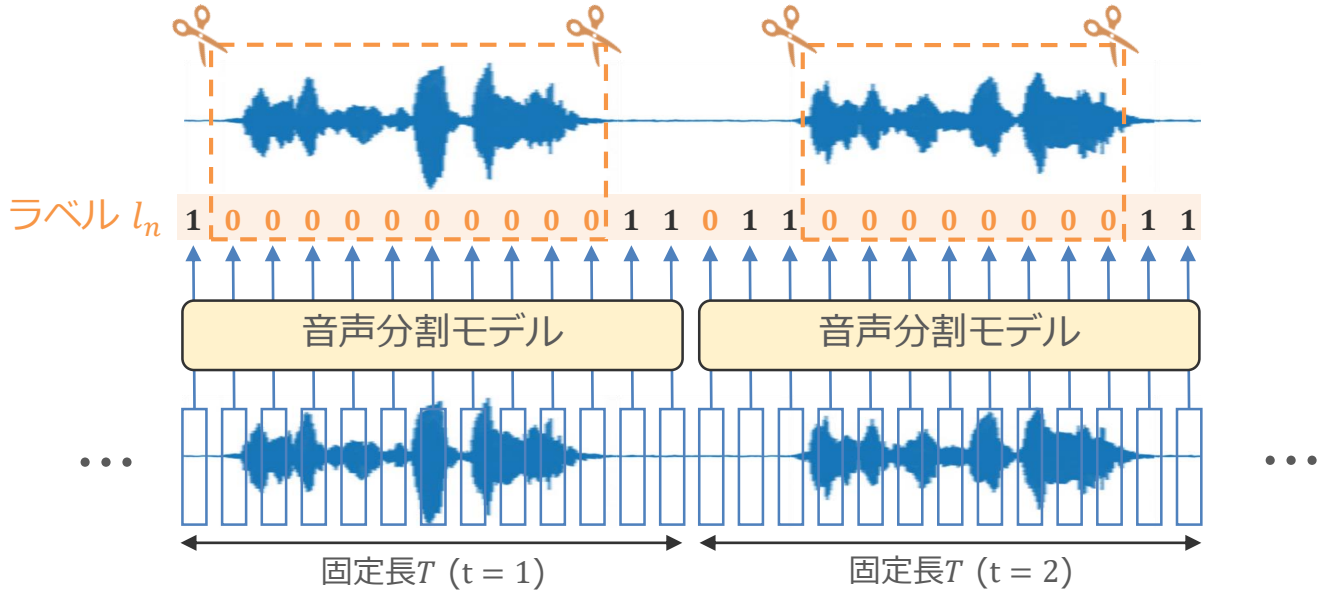

## モデルの学習

- 連続するセグメントを連結し、畳み込んだ音響特徴量にラベル付け
  - ラベル $x \in \{0,1\}$、ただし0はセグメント内、1はセグメント外のフレーム
- Binary Cross Entropy損失を最小化



# 提案手法：音声分割モデルの予測

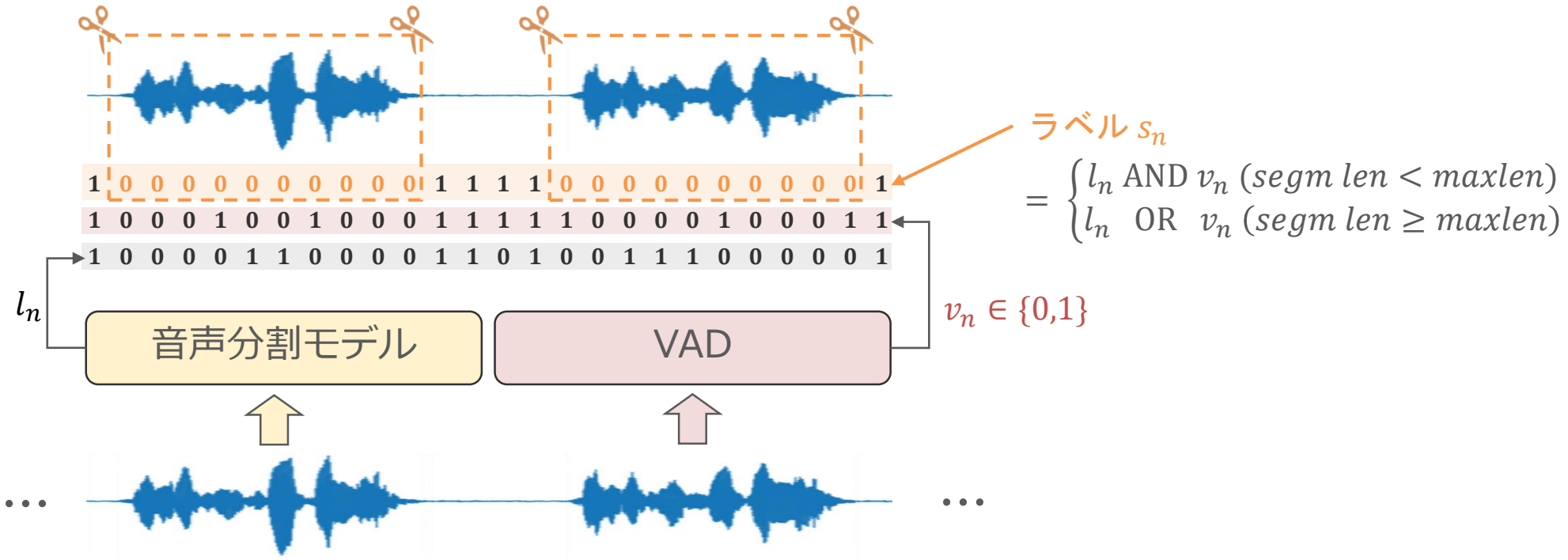
1. 固定長$T = 20$ (s)で分割した音声を音声分割モデルに入力
2. 各フレームのラベルを予測し、1の連続をセグメント境界として再分割
   - セグメントの長さが設定値以上になるように制約



# 提案手法：音声分割モデルとVADの組み合わせ

VADで検出した音声区間を用いて音声分割モデルの予測を補強

- 音声分割モデルの出力フレーム単位に変換したVADの予測を用いる
- 無音をセグメント境界の必要条件とすることで過剰分割を緩和



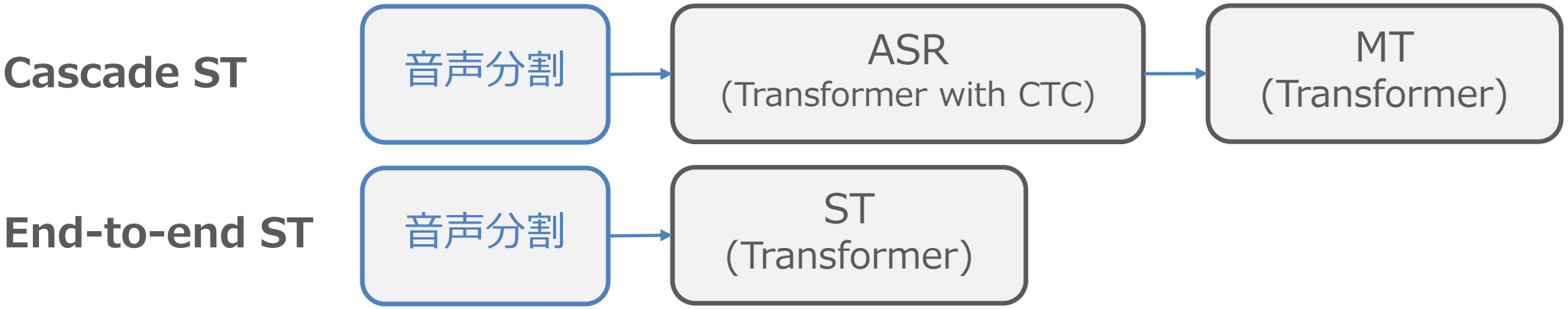
# 実験設定：データ、音声分割手法

- **データ**：MuST-C [Gangi+19] の英語→ドイツ語、英語→日本語
  - TED talksの音声翻訳コーパス
  - 音声分割モデルやSTの学習、および評価に使用

- **音声分割手法**
  - トップライン：人手による文単位の分割 (**人手分割**)
  - 比較手法：音声区間検出 (**VAD**)、固定長での分割 (**固定長**)
  - 提案手法：**音声分割モデル (+VAD)**

# 実験設定：STシステム、評価尺度

- **STシステム：**Cascade ST および End-to-end ST




- **評価尺度**
  - 機械翻訳（MT/ST）：BLEU
  - 音声認識（ASR）：Word Error Rate (WER)

# 事例分析① 出力テキストと分割位置

音声認識と機械翻訳出力（■はおよその分割位置）

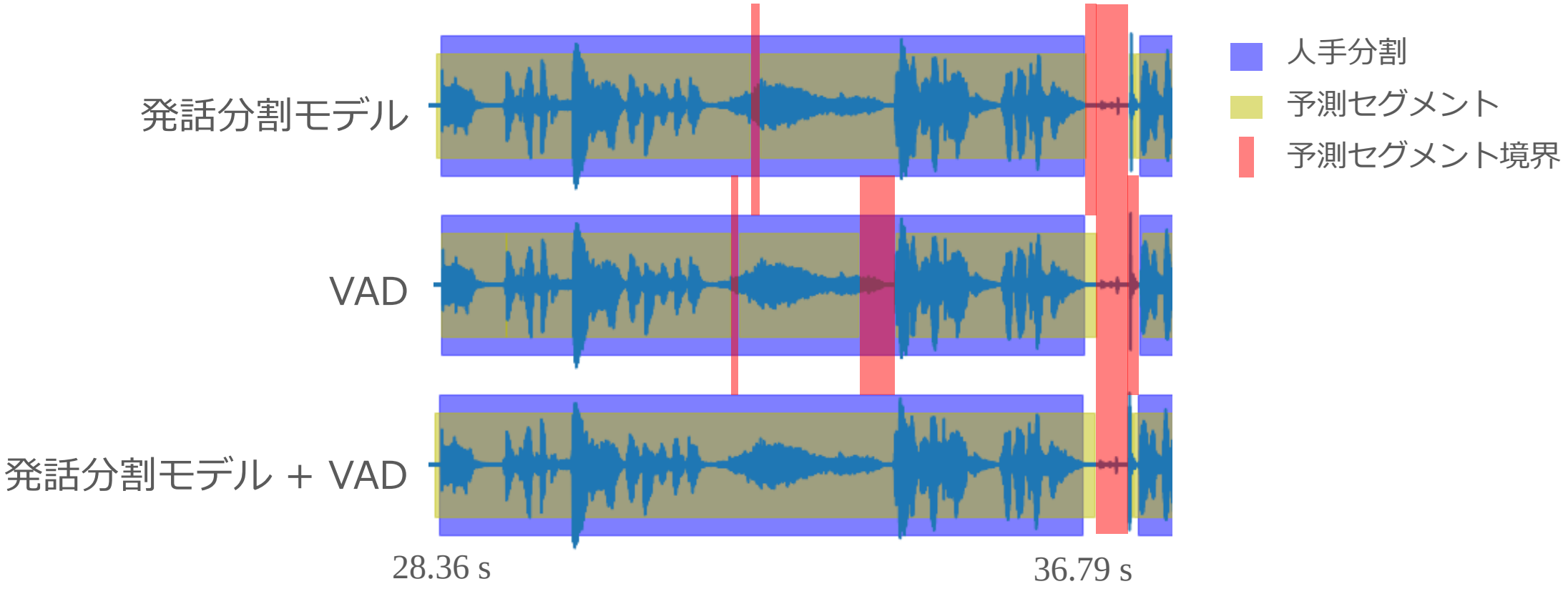
| | |
|---|---|
| 人手分割 (音声認識) | *bonobos are together with chimpanzees you aposre living closest relative* ■ |
| 人手分割 (機械翻訳) | *Bonobos sind zusammen mit Schimpansen, Sie leben am nächsten Verwandten.* ■ |
| VAD (音声認識) | *bonobos are* ■ *together with chimpanzees you aposre living closest relative that ...* |
| VAD (機械翻訳) | *Bonobos sind es.* ■ *Zusammen mit Schimpansen leben Sie im Verhältnis zum ...* |
| 発話分割モデル (音声認識) | *bonobos are together with chimpanzees you aposre living closest relative* ■ |
| 発話分割モデル (機械翻訳) | *Bonobos sind zusammen mit Schimpansen, Sie leben am nächsten Verwandten.* ■ |

- VADによる誤った分割 (“*bonobos are* ■ *together*”) と分割の見過ごし (“*relative that ...*”)
  → 翻訳結果に悪影響
- 提案手法は人手分割に近い位置で分割 → 人手分割と同じ翻訳結果

# 事例分析② 音声波形と分割位置

提案手法の発話分割モデルのみ、VADのみでは過剰な分割が目立った

→ 2つの予測を組み合わせることで過剰分割が緩和



# 発表内容

1. Speech Segmentation for Long-form Speech Translation
   - ① 音声分割 に関する取り組み

2. Streaming Simultaneous Speech Translation
   - ② 同時音声翻訳 に関する取り組み
   - ①と②を組み合わせたシステム構築

# 実験設定

実験1: 同時機械翻訳モデルの評価

- データ: MuST-C（TED talks）の英語→ドイツ語、英語→日本語、英語→中国語
- 評価尺度:BLEU（翻訳精度）, Average lagging (AL)（遅延）
- チャンクサイズ: 200~1000ms（大きいほど遅延大）

実験2: ストリーミングSTシステムの評価

- データ: MuST-C（TED talks）の英語→ドイツ語
- 評価尺度: BLEU（翻訳精度）
- チャンクサイズ: 400~1200ms

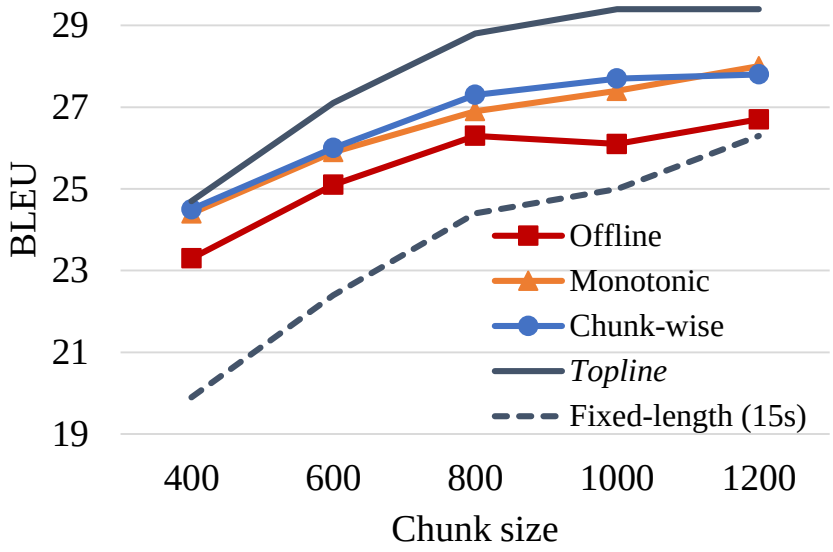
# Experiments: Streaming simultaneous ST

## Settings

- Dataset: MuST-C English-to-German
- Metrics: BLEU
- Latency control: chunk size = {400 to 1200 ms}

## Results

- Chunk-wise masking >= Monotonic masking > Offline model (unmasked)
- Segmentation models outperformed fixed-length baselines
- Chunk-wise masking retained over 94% of the segmentation in ST corpus (*Topline*)
  - More behind compared to 97% of offline settings (part 2)




| Chunk size | *Topline* | Fixed-length | Offline | Monotonic | Chunk-wise |
|---|---|---|---|---|---|
| 400 | 24.7 (100%) | 19.9 (80.6%) | 23.3 (94.3%) | 24.4 (98.8%) | 24.5 (99.2%) |
| 600 | 27.1 (100%) | 22.4 (82.7%) | 25.1 (92.6%) | 25.9 (95.6%) | 26 (95.9%) |
| 800 | 28.8 (100%) | 24.4 (84.7%) | 26.3 (91.3%) | 26.9 (93.4%) | 27.3 (94.8%) |
| 1000 | 29.4 (100%) | 25 (85.0%) | 26.1 (88.8%) | 27.4 (93.2%) | 27.7 (94.2%) |
| 1200 | 29.4 (100%) | 26.3 (89.5%) | 26.7 (90.8%) | 28 (95.2%) | 27.8 (94.6%) |